

Task-group Relatedness and Generalization Bounds for Regularized Multi-task Learning

Chao Zhang*, Dacheng Tao[†], Tao Hu[‡], Xiang Li[§]

August 29, 2014

Abstract

In this paper, we study the generalization performance of regularized multi-task learning (RMTL) in a vector-valued framework, where MTL is considered as a learning process for vector-valued functions. We are mainly concerned with two theoretical questions: 1) under what conditions does RMTL perform better with a smaller task sample size than STL? 2) under what conditions is RMTL generalizable and can guarantee the consistency of each task during simultaneous learning? In particular, we investigate two types of task-group relatedness: the observed discrepancy-dependence measure (ODDM) and the empirical discrepancy-dependence measure (EDDM), both of which detect the dependence between two groups of multiple related tasks (MRTs). We then introduce the Cartesian product-based uniform entropy number (CPUEN) to measure the complexities of vector-valued function classes. By applying the specific deviation and the symmetrization inequalities to the vector-valued framework, we obtain the generalization bound for RMTL, which is the upper bound of the joint probability of the event that there is at least one task with a large empirical discrepancy between the expected and empirical risks. Finally, we present a sufficient condition to guarantee the consistency of each task in the simultaneous learning process, and we discuss how task relatedness affects the generalization performance of RMTL. Our theoretical findings answer the aforementioned two questions.

Keywords: multi-task learning, generalization bound, task relatedness, consistency, vector-valued function

1 Introduction

There is plenty of empirical evidence to suggest that task-relatedness information improves multi-task learning (MTL) over single-task learning (STL) in multiple related task (MRT) scenarios.

*C. Zhang is with the School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, P.R. China. (e-mail: chao.zhang@dlut.edu.cn).

[†]D. Tao is with the Centre for Quantum Computation & Intelligent Systems, FEIT, University of Technology, Sydney, NSW 2007, Australia. (e-mail: dacheng.tao@gmail.com).

[‡]T. Hu is with the School of Mathematical Sciences, Capital Normal University, Beijing, 100048, P.R. China. (e-mail: hutaomath@foxmail.com).

[§]X. Li is with the School of Mathematical Sciences, Dalian University of Technology, Dalian, Liaoning, 116024, P.R. China. (e-mail: lixiangalixiang@gmail.com).

Therefore, capturing relatedness information is important for both theoretical and practical investigations of MTL.

Several learning methods have been proposed to address this problem. Evgeniou et al. [10] introduced regularized MTL to link the simultaneous learning process of MRT scenarios to STL problems, in which the regularization terms encode the relatedness between MRTs. However, regularization term design relies on a priori knowledge of tasks. Other methods that model task relatedness let the different tasks share common structures, *e.g.*, backpropagation networks [7] and the structure learning formulation [2]. Argyriou et al. [3] presented a method to learn a low-dimensional representation shared across MRTs, while Zhang and Yeung [23] applied covariance to model three types of relatedness between two tasks: the positive correlation, the negative correlation, and unrelatedness. From the theoretical standpoint, the notion “ \mathcal{F} -related” has been proposed to study the generalizability of multi-task classification, where if two tasks $\mathcal{Z}^{[1]}, \mathcal{Z}^{[2]}$ are \mathcal{F} -related for a given function class \mathcal{F} , there exists a function $f \in \mathcal{F}$ such that $P^{[1]} = f(P^{[2]})$ or $P^{[2]} = f(P^{[1]})$ [4, 5]. The interested reader is also referred to other theoretical investigations of MTL [16, 17] and learning theory [19, 8, 1, 24, 13, 12].

1.1 Overview of Main Results

As discussed by Micchelli and Pontil [20, 21], MTL can be studied from the viewpoint of vector-valued function learning. Inspired by [20, 21], we explore the vector-valued framework to study the generalization and consistency properties of regularized MTL (RMTL) and analyze the relationship between the properties of RMTL and task-group relatedness. In particular, we address the following theoretical questions:

- Under what conditions does RMTL perform better with a smaller task sample size than STL?
- Under what conditions is RMTL generalizable and can guarantee the consistency of each task during simultaneous learning?

In order to answer these questions, we also need to consider: 1) measures of task-group relatedness; 2) the joint probability of MRTs; 3) measures of vector-valued function classes; and 4) the specific deviation and symmetrization inequalities for the vector-valued framework.

Here, we introduce two types of task-group relatedness: the observed discrepancy-dependence measure (ODDM) and the empirical discrepancy-dependence measure (EDDM) (see Section 3).¹ ODDM measures the statistical dependence between events that some tasks have large observed discrepancies and the others have small observed discrepancies. EDDM measures the statistical dependence between events that some tasks have large empirical discrepancies and the others have small empirical discrepancies. In contrast to ODDM, EDDM reflects the asymptotic behavior of the relatedness between two task groups when the sample size goes to *infinity*.² We show that ODDM (or EDDM) can exist in three states: negative, positive, and *zero*, which respectively model three types of relatedness between two task groups: the synergy effect, the negative synergy effect, and unrelatedness.

¹In this paper, the observed discrepancy is defined as the discrepancy between an observation and its expectation, and the empirical discrepancy is defined as the discrepancy between the expectation (*i.e.*, expected risk) and its empirical estimate (*i.e.*, empirical risk).

²For convenience, we assume that all tasks have the same sample size in this paper.

Since MTL refers to a process in which MRTs are simultaneously processed, we consider the task joint probability, defined in (4), instead of the task summation probability as in [16, 17, 2, 11]. In task joint probability, the generalization bound for MTL is deemed to be the upper bound of the joint probability that there is at least one task with a large empirical discrepancy in MTL. This bound can also be used to describe the consistency of each task in the MTL learning process. In order to obtain the bound, we present the specific deviation inequalities and the symmetrization inequalities for the vector-valued framework and, meanwhile, introduce the Cartesian product-based uniform entropy number (CPUEN), which is induced from the uniform entropy numbers (UENs) of MRTs.

Based on the resulting generalization bounds, the theoretical properties of RMTL are analyzed and we show that:

- the validity of RMTL will theoretically be guaranteed if most of the relatedness between two task groups show a synergy effect. If almost any pair of task groups are predominantly mutual, RMTL performs well with less samples than STL, and the required sample size of each task in RMTL will not increase dramatically, regardless of the (large) number of MRTs (see Remarks 5.1&5.2).
- there will be a tighter generalization bound for RMTL if the values of EDDMs are negative, *i.e.*, if most of the relatedness between two task groups show a synergy effect. Moreover, we present a sufficient condition to guarantee the consistency of each task in RMTL.

Furthermore, we obtain the following theoretical findings:

- The aforementioned sufficient condition can be used to examine whether the given tasks, function classes, and regularization terms are suitable for MTL.
- The existence of a negative correlation between two tasks is necessary for MTL, which is in accordance with the argument by Zhang and Yeung [23].
- The generalization bound of RMTL.
- The relationship between the task relatedness and the generalization performance of RMTL.
- The sufficient condition to guarantee the consistency of each task in RMTL.
- The proposed vector-valued framework can be used to study the theoretical properties of vector-valued function learning [21]

1.2 Organization of the Paper

The rest of this paper is organized as follows. In Section 2, the main research addressed in this paper, including the task-joint probability and generalization bounds for RMTL, is formalized. In Section 3, two quantities for measuring task-group relatedness are presented and CPUEN is introduced in Section 4 to measure the complexity of the vector-valued function classes. The main results are presented in Section 5, along with a method to examine the validity of MTL. In Section 6, we address the generalization performance results using the covariance information of MRTs and the last section concludes the paper. In Appendix, we first present the deviation inequalities and the symmetrization inequalities for the vector-valued framework (Parts A & B). Finally, the proofs of the main results are given in Part C.

2 Problem Setup

We first formalize the main research addressed in this paper, including the task-joint probability and generalization bounds for RMTL.

2.1 Regularized Multi-task Learning

Given a space $\mathcal{X} \subset \mathbb{R}^I$, let $\mathcal{X}^{[m]}$ be the input space of the m -th task with the probability distribution $\mathcal{D}^{[m]}$ on \mathcal{X} and $\mathcal{Y}^{[m]} \in \mathbb{R}^J$ be the corresponding output space ($1 \leq m \leq M$). Let $g_*^{[m]} : \mathcal{X}^{[m]} \rightarrow \mathcal{Y}^{[m]}$ be the corresponding labeling function. Also, denote the m -th task as $\mathcal{Z}^{[m]} := \mathcal{X}^{[m]} \times \mathcal{Y}^{[m]} \subset \mathbb{R}^K$ with $K = I + J$.

In MTL, let $\mathcal{G}^{[1]}, \dots, \mathcal{G}^{[M]} \subset \mathcal{Y}^{\mathcal{X}}$ be M function classes corresponding to the learning tasks $\mathcal{Z}^{[1]}, \dots, \mathcal{Z}^{[M]}$, respectively. MTL is expected to simultaneously find M functions $\tilde{g}^{[1]}, \dots, \tilde{g}^{[M]}$ from $\mathcal{G}^{[1]}, \dots, \mathcal{G}^{[M]}$ such that each $\tilde{g}^{[m]}$ can minimize the expected risk of the corresponding task $\mathcal{Z}^{[m]}$ over $\mathcal{G}^{[m]}$:

$$\mathbb{E}^{[m]}(\ell^{[m]} \circ g^{[m]}) = \int \ell^{[m]}(g^{[m]}(\mathbf{x}^{[m]}), \mathbf{y}^{[m]}) dP^{[m]}(\mathbf{z}^{[m]}), \quad 1 \leq m \leq M, \quad (1)$$

where $\ell^{[m]}$ and $P^{[m]}(\mathbf{z}^{[m]})$ are the loss function and the probability distribution of the task $\mathcal{Z}^{[m]}$, respectively, with $\mathbf{z}^{[m]} := (\mathbf{x}^{[m]}, \mathbf{y}^{[m]})^T$.

Since the task distributions $P^{[1]}, \dots, P^{[M]}$ are usually unknown, the target functions $\tilde{g}^{[1]}, \dots, \tilde{g}^{[M]}$ cannot be directly obtained by minimizing the expected risks (1) of MRTs. Instead, the empirical risk minimization (ERM) principle can be used to handle this issue. For each task $\mathcal{Z}^{[m]}$, let $\mathbf{Z}_N^{[m]} := \{\mathbf{z}_n^{[m]}\}_{n=1}^N$ be a set of N i.i.d. samples drawn from $\mathcal{Z}^{[m]}$ with $\mathbf{z}_n^{[m]} := (\mathbf{x}_n^{[m]}, \mathbf{y}_n^{[m]})^T$. The following is the objective function of RMTL:

$$\sum_{m=1}^M \mathbb{E}_N^{[m]}(\ell^{[m]} \circ g^{[m]}) + rR(g^{[1]}, \dots, g^{[M]}),$$

where

$$\mathbb{E}_N^{[m]}(\ell^{[m]} \circ g^{[m]}) := \frac{1}{N} \sum_{n=1}^N \ell^{[m]}(g(\mathbf{x}_n^{[m]}), \mathbf{y}_n^{[m]}), \quad (2)$$

is the empirical risk of the task $\mathcal{Z}^{[m]}$, $R(g^{[1]}, \dots, g^{[M]})$ is the regularization term that is designed to encode the relatedness information between MRTs and $r > 0$ is the regularization parameter.

Alternatively, and as mentioned by Kakade et al. [14], the above regularized optimization can be equivalently rewritten as

$$\min_{R(g^{[1]}, \dots, g^{[M]}) \leq c} \sum_{m=1}^M \mathbb{E}_N^{[m]}(\ell^{[m]} \circ g^{[m]}),$$

where, instead of exploiting the regularization, a hard restriction $R(g^{[1]}, \dots, g^{[M]}) \leq c$ is set to combine the function classes $\mathcal{G}^{[1]}, \dots, \mathcal{G}^{[M]}$, which shrinks the original search space \mathcal{G} to \mathcal{G}_c^R .³

³For example, if $g^{[m]}(\mathbf{x}^{[m]}) = x^{[m]}$ for any $1 \leq m \leq M$, the original search space \mathcal{G} is the M -dimensional real space \mathbb{R}^M . Then, by setting the restriction $\sum_{m=1}^M (x^{[m]})^2 \leq c^2$, the original space \mathcal{G} will become an M -dimensional sphere \mathcal{G}_c^R with radius c .

Therefore, a proper regularization term $R(\mathbf{g})$ can correctly encode the relatedness between MRTs, reduce the computational cost, and improve the generalization performance. However, this design relies on a prior knowledge of the MRTs.

From the vector-valued function learning perspective [20, 21], RMTL aims to find a vector-valued function $\mathbf{g}_N = (g_N^{[1]}, \dots, g_N^{[M]})^T$ by simultaneously solving the M optimization problems:

$$\min_{\mathbf{g} \in \mathcal{G}_c^R} \left\{ E_N^{[m]}(\ell^{[m]} \circ g^{[m]}), \quad 1 \leq m \leq M \right\}, \quad (3)$$

where $\min_{\mathbf{g} \in \mathcal{G}_c^R}$ stands for a component-wise minimum operator defined in Section 2.2.

2.2 Notations of Vector Operations

For the discussion that follows, it is first necessary to describe some notations of vector operations. Given two vectors, $\mathbf{x} = (x^{[1]}, \dots, x^{[M]})^T$ and $\mathbf{y} = (y^{[1]}, \dots, y^{[M]})^T$, let $|\mathbf{x}| := (|x^{[1]}|, \dots, |x^{[M]}|)^T$ and denote the expression $\mathbf{x} > \mathbf{y}$ (resp. $\mathbf{x} \geq \mathbf{y}$) as $x^{[m]} > y^{[m]}$ (resp. $x^{[m]} \geq y^{[m]}$) for any $1 \leq m \leq M$. Similarly, we denote $\mathbf{x} < \mathbf{y}$ (resp. $\mathbf{x} \leq \mathbf{y}$) as $x^{[m]} < y^{[m]}$ (resp. $x^{[m]} \leq y^{[m]}$) for any $1 \leq m \leq M$.

Furthermore, given $(a^{[1]}, \dots, a^{[M]})^T \in \mathbb{R}^M$, we define the component-wise supremum operator

$$\sup_{\mathbf{g} \in \mathcal{G}} \{ (g^{[1]}(a^{[1]}), \dots, g^{[M]}(a^{[M]}))^T \}$$

with $\mathbf{g} = (g^{[1]}, \dots, g^{[M]})^T$ as follows: if the vector-valued function $\mathbf{g}_\dagger = (g_\dagger^{[1]}, \dots, g_\dagger^{[M]})^T$ achieves the supremum over \mathcal{G} , each component $g_\dagger^{[m]}$ of the vector \mathbf{g}_\dagger achieves the supremum $\sup_{g^{[m]} \in \mathcal{G}^{[m]}} \{g^{[m]}(a^{[m]})\}$ over $\mathcal{G}^{[m]}$. Similarly, we define the component-wise minimum operator as

$$\min_{\mathbf{g} \in \mathcal{G}} \{ (g^{[1]}(a^{[1]}), \dots, g^{[M]}(a^{[M]}))^T \}.$$

2.3 Task-joint Probability and Generalization Bounds

In general, the generalization bounds for STL refer to the upper bounds of the supremum

$$\sup_{g \in \mathcal{G}} |E(\ell \circ g) - E_N(\ell \circ g)|$$

with an alternative probability expression

$$\Pr \left\{ \sup_{g \in \mathcal{G}} |E(\ell \circ g) - E_N(\ell \circ g)| > \xi \right\},$$

whose upper bound describes the rarity of the event that the *empirical discrepancy* between the expected risk $E(\ell \circ g)$ and the empirical risk $E_N(\ell \circ g)$ is larger than a given positive constant ξ .

Since MRTs are processed simultaneously in MTL, the following task-joint probability is straightforward: for any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$,

$$\Pr \left\{ \sup_{\mathbf{g} \in \mathcal{G}_c^R} \left\{ \begin{pmatrix} |E^{[1]}(\ell^{[1]} \circ g^{[1]}) - E_N^{[1]}(\ell^{[1]} \circ g^{[1]})| \\ \vdots \\ |E^{[M]}(\ell^{[M]} \circ g^{[M]}) - E_N^{[M]}(\ell^{[M]} \circ g^{[M]})| \end{pmatrix} \right\} \preceq \begin{pmatrix} \xi^{[1]} \\ \vdots \\ \xi^{[M]} \end{pmatrix} \right\}, \quad (4)$$

which describes the rarity of the event in RMTL that there is at least one task $\mathcal{Z}^{[m]}$ with empirical discrepancy larger than the constant $\xi^{[m]}$. The upper bound of (4) is the so-called “generalization bound” for RMTL. Compared to the STL bound, the RMTL bound (4) not only reflects the generalization performance of each task, but also the dependence between simultaneously learned tasks, *i.e.*, how the success (or failure) of some tasks affects the performance of the others.

For convenience, we further define the loss function class:

$$\mathcal{F}^{[m]} := \{\mathbf{z}^{[m]} \mapsto \ell^{[m]}(g^{[m]}(\mathbf{x}^{[m]}), \mathbf{y}^{[m]}) : g^{[m]} \in \mathcal{G}^{[m]}\}, \quad 1 \leq m \leq M; \quad (5)$$

the Cartesian product $\mathcal{F} := \mathcal{F}^{[1]} \times \dots \times \mathcal{F}^{[M]}$ is called the “vector-valued function class” in the rest of this paper. Similarly, based on the regularized vector-valued function class $\mathcal{G}_c^{\mathbf{R}}$, we define the regularized loss vector-valued function class by

$$\mathcal{F}_c^{\mathbf{R}} := \{(\ell^{[1]} \circ g^{[1]}, \dots, \ell^{[M]} \circ g^{[M]})^T : (g^{[1]}, \dots, g^{[M]})^T \in \mathcal{G}_c^{\mathbf{R}}\}, \quad (6)$$

which is also termed the regularized vector-valued function class in the remainder of this paper. Briefly, we denote for any $\mathbf{f} := (f^{[1]}, \dots, f^{[M]})^T \in \mathcal{F}$,

$$\mathbb{E}^{[m]} f^{[m]} := \int f^{[m]}(\mathbf{z}^{[m]}) d\mathbb{P}^{[m]}(\mathbf{z}^{[m]}) \quad ; \quad \mathbb{E}_N^{[m]} f^{[m]} := \frac{1}{N} \sum_{n=1}^N f^{[m]}(\mathbf{z}_n^{[m]}), \quad (7)$$

and the generalization bound (4) is equivalently rewritten as $\Pr\left\{\sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} \{|\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}|\} \not\leq \xi\right\}$ with

$$\mathbf{E}\mathbf{f} := (\mathbb{E}^{[1]} f^{[1]}, \dots, \mathbb{E}^{[M]} f^{[M]})^T$$

and

$$\mathbf{E}_N\mathbf{f} := (\mathbb{E}_N^{[1]} f^{[1]}, \dots, \mathbb{E}_N^{[M]} f^{[M]})^T.$$

3 Measures of Task-group Relatedness

Some existing works on task relatedness already describe the relationship between two individual tasks, for instance the \mathcal{F} -related [5, 4] notion and covariances [23]. In MTL, it is also necessary to consider the relationship between two task groups. Here, we present two measures of task-group relatedness: the observed discrepancy-dependence measure (ODDM) and the empirical discrepancy-dependence measure (EDDM).

3.1 ODDM

In probability theory, the dependence between two events \mathcal{A} and \mathcal{B} can be detected using the quantity $\Pr\{\mathcal{A}|\mathcal{B}\} - \Pr\{\mathcal{A}\}$, where \mathcal{A} and \mathcal{B} are positively dependent if the conditional probability $\Pr\{\mathcal{A}|\mathcal{B}\}$ of \mathcal{A} given \mathcal{B} is greater than the probability $\Pr\{\mathcal{A}\}$ (*i.e.*, $\Pr\{\mathcal{A}|\mathcal{B}\} - \Pr\{\mathcal{A}\} > 0$), and they are negatively dependent if the inequality is reversed [6, 22]. We introduce ODDM and EDDM to measure the relatedness between two task groups in MTL, based on the quantity.

Definition 3.1 Given M tasks $\mathcal{Z}^{[1]}, \dots, \mathcal{Z}^{[M]}$ and a regularized vector-valued function class \mathcal{F}_c^R , let $\Lambda := \{1, \dots, M\}$ be an index set and $\Lambda^{[m]}$ be a subset of Λ with the cardinality of m . For any $\Lambda^{[m]} \subset \Lambda$ and any $\xi = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$, ODDM is defined as

$$\phi_{\mathcal{F}}(\Lambda^{[m]}, \xi) := \sup_{\mathbf{f} \in \mathcal{F}_c^R} \left\{ \Pr\{\mathcal{A}_{\Lambda^{[m]}} | \mathcal{B}_{\Lambda^{[m]}}\} - \Pr\{\mathcal{A}_{\Lambda^{[m]}}\} \right\},$$

where $\mathbf{f} = (f^{[1]}, \dots, f^{[M]})^T$, $\overline{\Lambda^{[m]}}$ stands for the complementary set of $\Lambda^{[m]}$ with $\Lambda^{[m]} \cup \overline{\Lambda^{[m]}} = \Lambda$, and the events $\mathcal{A}_{\Lambda^{[m]}} := \{s^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}}$ and $\mathcal{B}_{\Lambda^{[m]}} := \{s^{[i]} \leq \xi^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ w.r.t. the observed discrepancy

$$s^{[i]} := |E^{[i]} f^{[i]} - f^{[i]}(\mathbf{z}^{[i]})|$$

of the task $\mathcal{Z}^{[m]}$.

As defined above, ODDM measures the dependence between the events that the tasks in group $\Lambda^{[m]}$ have large observed discrepancies and the tasks in $\overline{\Lambda^{[m]}}$ have small observed discrepancies. In fact, ODDM is determined by the inherent characteristics of MRTs, the selection of function classes and the regularization term. It can exist in one of three states:

- a positive ODDM implies that some functions in the search space \mathcal{F}_c^R will result in a negative synergy effect between the tasks $\{\mathcal{Z}^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ and the others $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$, i.e., the success of tasks $\{\mathcal{Z}^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ will benefit from a performance loss in the others $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$;
- a negative ODDM means that all functions in \mathcal{F}_c^R will effect the synergy effect on the simultaneous learning process for MRTs, i.e., the success of the tasks $\{\mathcal{Z}^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ contributes to improved performance of the others $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$;
- a zero ODDM reflects that some functions in \mathcal{F}_c^R eliminate the relatedness between $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$ and $\{\mathcal{Z}^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$, and the others will effect synergy effect between the two groups.

3.2 EDDM

Since this paper focuses on ERM-based RM TL, we also need to consider the asymptotic behavior of the dependence between two task groups when the sample size N goes to *infinity*.

Definition 3.2 Following the notations in Definition 3.1 and letting $\mathbf{Z}_N^{[m]} := \{\mathbf{z}_n^{[m]}\}_{n=1}^N$ be N i.i.d. samples drawn from each task $\mathcal{Z}^{[m]}$ ($1 \leq m \leq M$), EDDM is defined as

$$\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \xi) := \Pr\{\mathcal{A}_{\Lambda^{[m]}}^N | \mathcal{B}_{\Lambda^{[m]}}^N\} - \Pr\{\mathcal{A}_{\Lambda^{[m]}}^N\},$$

where the events $\mathcal{A}_{\Lambda^{[m]}}^N := \{t_N^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}}$ and $\mathcal{B}_{\Lambda^{[m]}}^N := \{t_N^{[i]} \leq \xi^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ with the empirical discrepancy

$$t_N^{[i]} := \sup_{f \in \text{Prj}^{[i]}(\mathcal{F}_c^R)} |E^{[i]} f - E_N^{[i]} f|, \quad (8)$$

w.r.t. the sample set $\mathbf{Z}_N^{[m]}$ drawn from $\mathcal{Z}^{[m]}$, and $\text{Prj}^{[i]}(\mathcal{F}_c^R)$ stands for the projection of the regularized vector-valued function class \mathcal{F}_c^R onto the function class $\mathcal{F}^{[i]}$.

Note that EDDM measures the dependence between the generalization performances of the two task groups and also has three states:

- a positive EDDM implies that the successfully learned tasks $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$ benefit from a loss in generalization performance of the others $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$;
- a negative EDDM means that the task groups $\{\mathcal{Z}^{[i]}\}_{i \in \Lambda^{[m]}}$ and $\{\mathcal{Z}^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$ are mutually beneficial;
- a zero EDDM with $N < \infty$ signifies that the two groups are unrelated.

3.3 Empirically Computing ODDM and EDDM

By the facts that $\Pr\{\mathcal{A}|\mathcal{B}\} = \Pr\{\mathcal{A}, \mathcal{B}\}/\Pr\{\mathcal{B}\}$ and $\Pr\{\mathcal{A}\} = \mathbb{E}\mathbf{1}_{\{\mathcal{A}\}}$, ODDM $\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi})$ can be empirically computed in the following way. Letting $\{\mathbf{z}_n^{[m]}\}_{n=1}^N$ be i.i.d. samples drawn from the task $\mathcal{Z}^{[m]}$ ($1 \leq m \leq M$), we denote ζ_j ($1 \leq j \leq J$), η_k ($1 \leq k \leq K$) and θ_p ($1 \leq p \leq P$) as the observations of the events $\mathcal{A}_{\Lambda^{[m]}} \wedge \mathcal{B}_{\Lambda^{[m]}}$, $\mathcal{A}_{\Lambda^{[m]}}$ and $\mathcal{B}_{\Lambda^{[m]}}$, respectively. Then, an empirical version of ODDM $\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi})$ is given by:

$$\hat{\phi}_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) := \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} \left\{ \frac{J^{-1} \sum_{j=1}^J \mathbf{1}_{\{\zeta_j\}}}{P^{-1} \sum_{p=1}^P \mathbf{1}_{\{\theta_p\}}} - K^{-1} \sum_{k=1}^K \mathbf{1}_{\{\eta_k\}} \right\}, \quad (9)$$

where the expected risk $\mathbb{E}^{[i]} f^{[i]}$ in $s^{[i]}$ is approximated by its empirical version $\mathbb{E}_N^{[i]} f^{[i]}$.

Recalling the term $t_N^{[i]}$ defined in (8), EDDM $\varphi_{\mathcal{F}_c^{\mathbf{R}}}^N(\Lambda^{[m]}, \boldsymbol{\xi})$ can be approximately computed in the following way. First, fix the sample set $\{\mathbf{z}_n^{[i]}\}_{n=1}^N$ of each task $\mathcal{Z}^{[i]}$ ($1 \leq i \leq M$) and replace the expected risk $\mathbb{E}^{[i]} f$ with the fixed empirical quantity $\mathbb{E}_N^{[i]} f$ w.r.t. $\{\mathbf{z}_n^{[i]}\}_{n=1}^N$. Next, we randomly select L samples from of each task $\mathcal{Z}^{[i]}$ to form another empirical risk $\mathbb{E}_L^{[i]} f$ and denote $\hat{t}_L^{[i]} := \sup_{f \in \text{Prj}^{[i]}(\mathcal{F}_c^{\mathbf{R}})} |\mathbb{E}_L^{[i]} f - \mathbb{E}_N^{[i]} f|$ as an estimate of $t_N^{[i]}$. Denote the events $\mathcal{A}_L := \{\hat{t}_L^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}}$ and $\mathcal{B}_L := \{\hat{t}_L^{[i]} \leq \xi^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}$. Let ζ_j ($1 \leq j \leq J$), η_k ($1 \leq k \leq K$) and θ_p ($1 \leq p \leq P$) be the observations of the events $\mathcal{A}_L \wedge \mathcal{B}_L$, \mathcal{A}_L and \mathcal{B}_L respectively. We then can empirically compute EDDM $\varphi_{\mathcal{F}_c^{\mathbf{R}}}^N(\Lambda^{[m]}, \boldsymbol{\xi})$ as

$$\hat{\varphi}_{\mathcal{F}_c^{\mathbf{R}}}^N(\Lambda^{[m]}, \boldsymbol{\xi}) := \frac{J^{-1} \sum_{j=1}^J \mathbf{1}_{\{\zeta_j\}}}{P^{-1} \sum_{p=1}^P \mathbf{1}_{\{\theta_p\}}} - K^{-1} \sum_{k=1}^K \mathbf{1}_{\{\eta_k\}}. \quad (10)$$

Remark 3.1 *There are two difficulties to implement this method to empirically compute ODDM and EDDM:*

- *In general, it is hard to capture the observations of the task-joint events.*
- *If the task number M is large, it is highly time-consuming to compute the empirical estimates of ODDM and EDDM for any $\Lambda^{[m]}$. To reduce the complexity, one feasible way is to cluster the tasks according to the similarity and select a representative task from each cluster to compute ODDM and EDDM.*

4 Cartesian Product-based Uniform Entropy Numbers

Complexity measures of function classes play an important role in learning theory. Since this paper studies MTL in the vector-valued framework, the classical measures such as the Vapnik-Chervonenkis (VC) dimension and the covering number, are not applicable (or at least cannot be directly applied) to the vector-valued scenario. For example, Ben-David and Borbely [4] applied an extended version of the VC dimension to study the generalization properties of multi-task classification.

Here, we introduce the Cartesian product-based uniform entropy number (CPUEN) to measure the complexity of the vector-valued function classes. First, we briefly outline the definitions of the covering number and uniform entropy number (UEN) of the scalar-valued function classes. Regarding further details, please refer to Mendelson [18].

Definition 4.1 *Let \mathcal{F} be a function class and d be a metric on \mathcal{F} . For any $\xi > 0$, the covering number of \mathcal{F} at radius ξ w.r.t. the metric d , denoted by $\mathcal{N}(\mathcal{F}, \xi, d)$ is the minimum size of a cover of radius ξ . Furthermore, given a sample set $\mathbf{Z}_N := \{\mathbf{z}_n\}_{n=1}^N$ drawn from \mathcal{Z} , we denote $\mathbf{Z}'_N := \{\mathbf{z}'_n\}_{n=1}^N$ as the ghost sample set drawn from \mathcal{Z} , such that the ghost sample \mathbf{z}'_n has the same distribution as \mathbf{z}_n for any $1 \leq n \leq N$. Denote $\mathbf{Z}_{2N} := \{\mathbf{Z}_N, \mathbf{Z}'_N\}$. Setting the metric d as the $\ell_p(\mathbf{Z}_{2N})$ ($p > 0$) norm, UEN is defined by*

$$\ln \mathcal{N}_p(\mathcal{F}, \xi, N) := \sup_{\mathbf{Z}_N} \ln \mathcal{N}(\mathcal{F}, \xi, \ell_p(\mathbf{Z}_N)). \quad (11)$$

Recall that the vector-valued function class \mathcal{F} is a Cartesian product of the function classes $\mathcal{F}^{[1]}, \dots, \mathcal{F}^{[M]}$, i.e., $\mathcal{F} := \mathcal{F}^{[1]} \times \dots \times \mathcal{F}^{[M]}$. For each $\mathcal{F}^{[m]}$ ($1 \leq m \leq M$), let $\tilde{\mathbf{Z}}_N^{[m]}$ be the sample set achieving the supremum

$$\sup_{\mathbf{Z}_N^{[m]} \in (\mathcal{Z}^{[m]})^N} \ln \mathcal{N}(\mathcal{F}^{[m]}, \xi^{[m]}, \ell_p(\mathbf{Z}_N^{[m]})) \quad (12)$$

and $\Omega_{p,N}^{[m]}$ be one of the covers of $\mathcal{F}^{[m]}$ related to the supremum w.r.t. the norm $\ell_p(\tilde{\mathbf{Z}}_N^{[m]})$. Therefore, the Cartesian product $\Omega_{p,N}^{[1]} \times \dots \times \Omega_{p,N}^{[M]}$ is also a cover of \mathcal{F} with the radius vector $\boldsymbol{\xi} := (\xi^{[1]}, \dots, \xi^{[M]})^T$. Following the above notations, we define the CPUEN of the vector-valued function class \mathcal{F} as follows:

Definition 4.2 *Given a vector-valued function class \mathcal{F} , consider a Cartesian product-based cover of the vector-valued function \mathcal{F} :*

$$\Omega_{p,N}(\mathcal{F}, \boldsymbol{\xi}) := \left\{ \mathcal{A}_p^M \in \Omega_{p,N}^{[1]} \times \dots \times \Omega_{p,N}^{[M]} : \mathcal{A}_p^M \cap \mathcal{F} \neq \emptyset \right\}.$$

Then, CPUEN of \mathcal{F} is defined as $\ln \mathcal{N}_p(\mathcal{F}, \boldsymbol{\xi}, N) := \ln |\Omega_{p,N}(\mathcal{F}, \boldsymbol{\xi})|$.

In contrast to the classical UEN [see (11)], CPUEN is induced from the cover of the function class $\mathcal{F}^{[m]}$ of each task $\mathcal{Z}^{[m]}$ ($1 \leq m \leq M$) with different norms and radiuses instead of introducing a uniform norm in the vector-valued function space \mathcal{F} . Although CPUEN is usually larger than the uniform-norm UEN of the vector-valued function class \mathcal{F} , the induction setting of CPUEN has a stronger relationship with the prior information-based design of the regularization term and offers convenience to the theoretical analysis of RMTL.

5 Generalization Bounds of Regularized Multi-task Learning

In this section, we present the generalization bounds of RMTL and discuss how the task-group relatedness affects the generalization properties of RMTL. Moreover, we give a sufficient condition for the consistency of each task in MRTs.

5.1 Two Special Cases

Before the formal discussion, we first bound the probabilities of two special events: first, that all tasks have large empirical discrepancies and second, that all tasks have small empirical discrepancies.

Theorem 5.1 *Assume that \mathcal{F}_c^R is a regularized vector-valued function class w.r.t. the constant c , and $\mathbf{Z}_N^{[m]} = \{\mathbf{z}_n^{[m]}\}_{n=1}^N$ is the set of N i.i.d. samples drawn from the task $\mathcal{Z}^{[m]}$ ($1 \leq m \leq M$). Let $\Lambda := \{1, \dots, M\}$ be an index set and denote $\Lambda^{[m]}$ as a subset of Λ with the cardinality of m . Denote $\mathbf{Z}_{2N}^{[m]} := \{\mathbf{Z}_N^{[m]}, \mathbf{Z}'_N^{[m]}\}$. Given $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$ and for any $N \in \mathbb{N}$ such that $N \geq \frac{8\Gamma(\Lambda)}{1-2\Upsilon(\Lambda)}$, it then holds that*

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2^{M+2} \mathcal{N}_1(\mathcal{F}_c^R, \boldsymbol{\xi}/8, 2N) \exp \left\{ \frac{-N \sum_{m=1}^M (\xi^{[m]})^2}{32M^2(b-a)^2} \right\}, \quad (13)$$

where

$$\Gamma(\Lambda) := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{m(b-a)^2}{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2}, \quad (14)$$

and

$$\Upsilon(\Lambda) := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}). \quad (15)$$

This theorem shows that if it holds that $N \geq \frac{8\Gamma(\Lambda)}{1-2\Upsilon(\Lambda)}$, the probability of $\sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}| > \boldsymbol{\xi}$ can be bounded by the RHS of (13). Note that if $M = 1$, since $\phi_{\mathcal{F}}(\Lambda^{[1]}, \boldsymbol{\xi})$ equals zero, the quantity $\Upsilon(\Lambda)$ is zero and the bound (13) coincides with the classical result of STL (see Theorem 2.3 of [18]).

Remark 5.1 *In the case of $M > 1$, the condition $N \geq \frac{8\Gamma(\Lambda)}{1-2\Upsilon(\Lambda)}$ should be satisfied when the quantity $\Upsilon(\Lambda) < 1/2$: namely, it is necessary for RMTL to satisfy the condition that the task-group relatedness between MRTs should mostly be synergistic. Furthermore, RMTL will perform well with less samples than STL size $N \geq \frac{8(b-a)^2}{\xi^2}$ if the condition $\Upsilon(\Lambda) \leq (1-2^{M-1})$ holds, which implies that almost any pair of task groups $\Lambda^{[m]}$ and $\overline{\Lambda}^{[m]}$ predominantly promote mutually.⁴*

⁴Actually, letting $\xi_0 := \min\{\xi^{[1]}, \dots, \xi^{[M]}\}$ and $N_0 := \frac{8(b-a)^2}{\xi_0^2}$, we have $8\Gamma(\Lambda) < (2^M - 1)N_0$. Thus, the condition $N \geq \frac{8\Gamma(\Lambda)}{1-2\Upsilon(\Lambda)}$ holds if N is larger than $\frac{(2^M - 1)N_0}{1-2\Upsilon(\Lambda)}$. We can then infer that each task in RMTL will need less samples than the task in STL if the condition $\frac{2^M - 1}{1-2\Upsilon(\Lambda)} < 1$ holds.

Remark 5.2 If $\xi = \xi^{[1]} = \dots = \xi^{[M]}$ and each ODDM $\phi_{\mathcal{F}}(\Lambda^{[m]}, \xi)$ reaches the minimum value -1 , the sample size N of each task should be larger than the value $\frac{8(b-a)^2}{((2^M-1)^{-1}+2)\xi^2}$ ($M > 1$) to support the inequality (13). This implies that the required sample size of each task in RMTL will approach half of the STL value $8(b-a)^2/\xi^2$ at the rate of 2^{-M} as $M \rightarrow \infty$. This finding shows that if the relationship between any pair of task groups $\Lambda^{[m]}$ and $\overline{\Lambda}^{[m]}$ is predominantly synergistic, each task in RMTL needs less samples than STL and the required sample size N in RMTL will not increase dramatically, regardless of a large number of MRTs.

We next consider the second special case and present an upper bound of the probability that all tasks have small empirical discrepancies in the simultaneous learning process for MRTs. The following theorem is proved by using the small-deviation techniques [15].

Theorem 5.2 Following the notations in Theorem 5.1, it then holds that for any $\xi = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$,

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| \leq \xi \right\} \leq 2^M \sup_{\mathbf{f} \in \mathcal{F}_c^R} \Pr \{ \mathbf{s} \leq 2\xi \}, \quad (16)$$

where $\mathbf{s} = (s^{[1]}, \dots, s^{[M]})^T$ with $s^{[m]} := |\mathbf{E}^{[m]} f^{[m]} - f^{[m]}(\mathbf{z}^{[m]})|$ for any $1 \leq m \leq M$.

This theorem converts the case of small empirical discrepancies into a simple case, where the LHS of (16) can be bounded by using the probability that the observed discrepancy of each task $\mathcal{Z}^{[m]}$ is smaller than $2\xi^{[m]}$ ($1 \leq m \leq M$). Compared to the case of empirical discrepancies, the RHS of (16) is only determined by the inherent characteristics of MRTs, *e.g.*, the distributions of tasks, the selection of function classes, and the regularization term.

5.2 Main Results

Based on these two special cases, we obtain the generalization bounds of RMTL and a sufficient condition for the consistency of each task in the simultaneous learning process for MRTs.

Theorem 5.3 Following the notations of Theorem 5.1, given $\xi = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$ and for any $N \in \mathbb{N}$ such that $N \geq \max_{1 \leq m \leq M} \max_{\Lambda^{[m]} \subset \Lambda} \frac{8\Gamma(\Lambda^{[m]})}{1-2\Upsilon(\Lambda^{[m]})}$, it then holds that

$$\begin{aligned} \Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| \not\leq \xi \right\} &\leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} 2^m \Pr \left\{ \{s^{[\lambda]} \leq 2\xi^{[\lambda]}\}_{\lambda \in \overline{\Lambda}^{[m]}} \right\} \\ &\times \left(\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \xi) + 2^{m+2} \mathcal{N}_1 \left(\text{Prj}_{\Lambda^{[m]}}^{\mathcal{F}_c^R}, \frac{\xi_{\Lambda^{[m]}}}{8}, 2N \right) \exp \left\{ \frac{-N \sum_{\lambda \in \Lambda^{[m]}} (\xi^{[\lambda]})^2}{32M^2(b-a)^2} \right\} \right), \end{aligned} \quad (17)$$

where $\text{Prj}_{\Lambda^{[m]}}^{\mathcal{F}_c^R}$ stands for the projection of \mathcal{F}_c^R on the subspace $\prod_{\lambda \in \Lambda^{[m]}} \mathcal{F}^{[\lambda]}$, $\xi_{\Lambda^{[m]}} := (\xi^{[\lambda]})_{\lambda \in \Lambda^{[m]}}$,

$\Gamma(\Lambda^{[m]})$ and $\Upsilon(\Lambda^{[m]})$ are defined in (14). Furthermore, if it is satisfied that for any $1 \leq m \leq M$ and $\lambda^{[m]} \subset \Lambda$,

$$\lim_{N \rightarrow +\infty} \varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \xi) = \lim_{N \rightarrow +\infty} \ln \mathcal{N}_1 \left(\text{Prj}_{\Lambda^{[m]}}^{\mathcal{F}_c^R}, \frac{\xi_{\Lambda^{[m]}}}{8}, 2N \right) = 0, \quad (18)$$

it then holds that

$$\lim_{N \rightarrow +\infty} \Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| \not\leq \boldsymbol{\xi} \right\} = 0. \quad (19)$$

In this theorem, we obtain an upper bound of the joint probability of the event that $\sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| \not\leq \boldsymbol{\xi}$ and show that the consistency of each task in MTL can be guaranteed if condition (18) is valid. We are concerned with two aspects of the theorem:

- the RHS of (17) implies that given $N < \infty$, a smaller value of EDDM $\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \boldsymbol{\xi})$ will lead to a sharper bound, which is in accordance with the argument that the negative EDDM means that the task groups benefit from each other (see Section 3).
- The asymptotic convergence of the generalization bound is determined by two factors: 1) EDDM $\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \boldsymbol{\xi})$; and 2) CPUEN $\ln \mathcal{N}_1(\text{Prj}_{\Lambda^{[m]}}, \boldsymbol{\xi}_{\Lambda^{[m]}}/8, 2N)$. In particular, according to the classical results of STL (see Theorem 2.3 & Definition 2.5 of [18]), if UEN for each task $\mathcal{Z}^{[m]}$ satisfies that $\frac{\ln \mathcal{N}_1(\mathcal{F}^{[m]}, \boldsymbol{\xi}^{[m]}/8, 2N)}{N}$ converges to zero when N goes to infinity, the second equality of (18) holds. Note that the convergence of $\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \boldsymbol{\xi})$ is determined by the inherent characteristics of MRTs, *e.g.*, distributions of tasks, selection of function classes, and regularization terms.

Remark 5.3 *Moreover, these theoretical findings cause us to preliminarily examine whether the combination of tasks, function classes, and regularization terms is suitable for the ERM-based RMTL according to the rules that*

$$\Upsilon(\Lambda) = \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) < \frac{1}{2},$$

and

$$\lim_{N \rightarrow +\infty} \varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \boldsymbol{\xi}) = 0$$

with $\varphi_{\mathcal{F}_c^R}^N(\Lambda^{[m]}, \boldsymbol{\xi}) \leq 0$ for any $\Lambda^{[m]} \subset \Lambda$ ($1 \leq m \leq M$).

6 Generalization Bounds with Covariance Information

As discussed in Section 3, since ODDM detects the dependence between two task groups, the bound (13) cannot reflect how the individual relatedness between two tasks affects the generalization performance of RMTL for more than two tasks. Here, we consider the generalization results based on the covariance information between every two tasks.

Theorem 6.1 *Follow the notations of Theorem 5.1. Given $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$ and for any $N \in \mathbb{N}$ such that*

$$N \geq \frac{8\Gamma_2}{1 - 2(\Upsilon(\Lambda) + \Upsilon_2)}, \quad (20)$$

then there holds that

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2^{M+2} \mathcal{N}_1(\mathcal{F}_c^R, \boldsymbol{\xi}/8, 2N) \exp \left\{ \frac{-N \sum_{m=1}^M (\xi^{[m]})^2}{32M^2(b-a)^2} \right\}, \quad (21)$$

where

$$\Gamma_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{m(b-a)^2}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2}, \quad (22)$$

and

$$\Upsilon_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{8 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}_{\mathcal{F}}(i_1, i_2)}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2}. \quad (23)$$

Compared to Theorem 5.1, the condition (20) incorporates the quantity Υ_2 which is related to the covariance information. Actually, the quantity Υ_2 is derived by replacing $\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2$ with $\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2$ as shown in the proof of Lemma B.2. From the condition (20), we can find that the bound (21) is valid when $\Upsilon(\Lambda) + \Upsilon_2 < 1/2$, which means that if the synergetic effect is the main group relatedness in the learning process and some of the correlations between tasks are negative, the learning process will perform well with a small sample size N . Zhang and Yeung [23] have highlighted the necessity of the negative correlation and pointed out that the negative correlation is helpful to reduce the search space in MTL, which is in accordance with our theoretical findings.

However, when $M = 1$, the bound (21) coincides with the canonical results in STL if and only if the quantity $\text{Cov}_{\mathcal{F}}(i, i)$ equals to *zero*, *i.e.*, the random variable \mathbf{z} of the task $\mathcal{Z}^{[i]}$ takes a constant with the probability of *one*. Since this setting is far away from the practical scenario, unlike the result (17), the bound (21) that encodes covariance information cannot reflect the transition from STL to MTL.

7 Conclusion

In this paper, we apply the vector-valued framework to study the generalization performance of RMTL and analyze the relationship between the task-group relatedness and the properties of RMTL. In particular, we introduce two types of task-group relatedness: ODDM and EDDM, and we present CPUEN to measure the complexity of the regularized vector-valued function class $\mathcal{F}_c^{\mathbf{R}}$. By applying the specific deviation and symmetrization inequalities to the vector-valued framework, we obtain the generalization bound for RMTL and provide a sufficient condition to guarantee the consistency of each task in the simultaneous learning process of MRTs. Finally, we show that the theoretical findings of this paper can examine whether the task settings are suitable for the RMTL mechanism.

Based on the theoretical findings, we summarize the relationship between the generalization properties of RMTL and the task-group relatedness as follows:

- ODDM is related to the sample size and validity of RMTL (see Theorem 5.1). We first prove that the condition of $\Upsilon(\Lambda) < \frac{1}{2}$ is necessary for the validity of RMTL and then show that if almost any pair of task groups $\Lambda^{[m]}$ and $\overline{\Lambda^{[m]}}$ predominantly mutually promote, the required sample size N of each task in RMTL will be smaller than that of STL for each

task. The sample size will also not increase dramatically, regardless of a large number of MRTs (see Remarks 5.1 & 5.2).

- EDDM affects the generalization performance of RMTL as follows: 1) a negative EDDM provides a sharper generalization bound; and 2) the asymptotic behavior of EDDM also affects the consistency of the task (see Theorem 5.3).
- The existence of a negative correlation between two tasks is necessary for MTL, which is in accordance with the relevant argument of [23].

In summary, synergistic task-group relatedness is beneficial to the generalization performance of RMTL. In future works, we will focus on the practical applications of the theoretical findings, for instance by improving the empirical computations of ODDM and EDDM (see Remark 5.3) and designing the regularization term for RMTL based on the task-group relatedness.

References

- [1] A. Agarwal and J.C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013.
- [2] R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.
- [3] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *Advances in neural information processing systems (NIPS)*, 19:41, 2007.
- [4] S. Ben-David and R.S. Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.
- [5] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Learning Theory and Kernel Machines*, pages 567–580. Springer, 2003.
- [6] R.C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.
- [7] R. Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.
- [8] N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Transactions on Information Theory*, 54(1):386–390, 2008.
- [9] X. Chen. Concentration inequalities for bounded random vectors. *arXiv preprint arXiv:1309.0003*, 2013.
- [10] T. Evgeniou, C. A. Micchelli, M. Pontil, and J. Shawe-Taylor. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6(4):615–637, 2005.
- [11] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

- [12] Z. Hussain, J. Shawe-Taylor, D.R. Hardoon, and C. Dhanjal. Design and generalization analysis of orthogonal matching pursuit algorithms. *IEEE Transactions on Information Theory*, 57(8):5326–5341, 2011.
- [13] R. Jin, T. Yang, M. Mahdavi, Y. Li, and Z. Zhou. Improved bounds for the nyström method with application to kernel classification. *IEEE Transactions on Information theory*, 59(10):6939–6949, 2013.
- [14] S.M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2008.
- [15] W.V. Li. Small value probabilities: Techniques and applications. *Lecture notes*, 2012.
- [16] A. Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7:117–139, 2006.
- [17] A. Maurer, M. Pontil, and B. Romera-Paredes. Sparse coding for multitask and transfer learning. In *Proceedings of The 30th International Conference on Machine Learning (ICML’13)*, pages 343–351, 2013.
- [18] S. Mendelson. A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*, pages 1–40, 2003.
- [19] S. Mendelson. Lower bounds for the empirical minimization algorithm. *IEEE Transactions on Information Theory*, 54(8):3797–3803, 2008.
- [20] C.A. Micchelli and M. Pontil. Kernels for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 921–928, 2004.
- [21] C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.
- [22] M. Mohri and A. Rostamizadeh. Stability bounds for stationary φ -mixing and β -mixing processes. *Journal of Machine Learning Research*, 11:789–814, 2010.
- [23] Y. Zhang and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI’10)*, pages 733–742, 2010.
- [24] D. Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.

A Deviation Inequalities for Random Vectors

To obtain the generalization bounds for RMTL, we need to consider the deviation inequalities for random vectors. The following lemma is derived from [9].

Let $\mathbf{s}_n = (s_n^{[1]}, \dots, s_n^{[M]})^T \in \mathbb{R}^M$ ($1 \leq n \leq N$) be N i.i.d. random vectors such that

$$\sum_{m=1}^M s_n^{[m]} \leq 1, \quad \text{for } n = 1, \dots, N, \quad (24)$$

and

$$s_n^{[m]} \geq 0, \quad \text{for } 1 \leq n \leq N \text{ and } 1 \leq m \leq M. \quad (25)$$

Note that the components $s_n^{[1]}, \dots, s_n^{[M]}$ of \mathbf{s}_n are not necessarily independent. The mean $\boldsymbol{\mu} = (\mu^{[1]}, \dots, \mu^{[M]})^T$ of random vectors \mathbf{s}_n is expressed as

$$\mu^{[m]} = \mathbb{E}^{[m]} s_n^{[m]}, \quad \text{for } 1 \leq m \leq M. \quad (26)$$

Lemma A.1 *For any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$ such that $\sum_{m=1}^M (\mu^{[m]} + \xi^{[m]}) < 1$, then there holds that*

$$\Pr \left\{ \left| \frac{1}{N} \sum_{n=1}^N \mathbf{s}_n - \boldsymbol{\mu} \right| > \boldsymbol{\xi} \right\} \leq 2^M \exp \left\{ -2N \sum_{m=1}^M (\xi^{[m]})^2 \right\}. \quad (27)$$

Moreover, since the vector-valued function \mathbf{f} has the range $[a, b]$, let

$$s_n^{[m]} := \frac{f^{[m]}(\mathbf{z}_n^{[m]}) - a}{M(b - a)}, \quad 1 \leq n \leq N, \quad 1 \leq m \leq M, \quad (28)$$

and then

$$\Pr \left\{ |\mathbf{E}_N \mathbf{f} - \mathbf{E} \mathbf{f}| > \boldsymbol{\xi} \right\} = \Pr \left\{ \left| \frac{1}{N} \sum_{n=1}^N \mathbf{s}_n - \frac{\mathbf{E} \mathbf{f} - \mathbf{a}}{M(b - a)} \right| > \frac{\boldsymbol{\xi}}{M(b - a)} \right\}, \quad (29)$$

where $\mathbf{a} = (a, \dots, a)^T \in \mathbb{R}^M$. Thus, the combination of Lemma A.1 and (29) leads to a Hoeffding-type deviation inequality for random vectors.

Theorem A.1 *Given a bounded vector-valued function $\mathbf{f} = (f^{[1]}, \dots, f^{[M]})^T$ with the range $[a, b]$, there holds that for any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$,*

$$\Pr \left\{ |\mathbf{E}_N \mathbf{f} - \mathbf{E} \mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2^M \exp \left\{ -2N \sum_{m=1}^M \frac{(\xi^{[m]})^2}{M^2(b - a)^2} \right\}. \quad (30)$$

B Symmetrization Inequalities for Random Vectors

B.1 Chebyshev Inequalities for Random Vectors

Definition B.1 *Assume that $\mathcal{Z}^{[1]}, \dots, \mathcal{Z}^{[M]}$ are M distributions on \mathbb{R} . Let $\boldsymbol{\Lambda} := \{1, \dots, M\}$ be an index set and $\Lambda^{[m]}$ be a subset of $\boldsymbol{\Lambda}$ with the cardinality of m . For any $\Lambda^{[m]} \subset \boldsymbol{\Lambda}$ and any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$, define*

$$\psi(\Lambda^{[m]}, \boldsymbol{\xi}) := \Pr \left\{ \{s^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}} \mid \{s^{[i]} \leq \xi^{[i]}\}_{i \in \overline{\Lambda^{[m]}}} \right\} - \Pr \left\{ \{s^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}} \right\}. \quad (31)$$

where $s^{[i]}$ is the non-negative random variable of the task $\mathcal{Z}^{[i]}$, and $\overline{\Lambda^{[m]}}$ stands for the complementary set of $\Lambda^{[m]}$ with $\Lambda^{[m]} \cup \overline{\Lambda^{[m]}} = \mathbf{\Lambda}$.

Lemma B.1 Let $\mathbf{s} = (s^{[1]}, \dots, s^{[M]})^T$ be a random vector with nonnegative elements and $\mathbf{\Lambda} = \{1, \dots, M\}$ be an index set. For any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$, then there holds that

$$\Pr \{\mathbf{s} \not\leq \boldsymbol{\xi}\} \leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\sum_{i \in \Lambda^{[m]}} \mathbb{E}^{[i]} \{(s^{[i]})^2\}}{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right), \quad (32)$$

where $\mathbf{s} \not\leq \boldsymbol{\xi}$ means that there is at least one index $m \in \mathbf{\Lambda}$ such that $s^{[m]} > \xi^{[m]}$, and $\Lambda^{[m]}$ stands for an index set with the cardinality of m .

Lemma B.2 Let $\mathbf{s} = (s^{[1]}, \dots, s^{[M]})^T$ be a random vector with nonnegative elements and $\mathbf{\Lambda} = \{1, \dots, M\}$ be an index set. For any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$, then there holds that

$$\Pr \{\mathbf{s} \not\leq \boldsymbol{\xi}\} \leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\sum_{i \in \Lambda^{[m]}} \mathbb{E}^{[i]} \{(s^{[i]})^2\} + 2 \sum_{i < j} \mathbb{E} \{s^{[i]} s^{[j]}\}}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right), \quad (33)$$

where $\mathbf{s} \not\leq \boldsymbol{\xi}$ means that there is at least one index $m \in \mathbf{\Lambda}$ such that $s^{[m]} > \xi^{[m]}$, and $\Lambda^{[m]}$ stands for an index set with the cardinality of m .

B.2 Symmetrization Inequalities

By applying ODDM, we can develop the symmetrization inequality for MTL as follows:

Theorem B.1 Assume that \mathcal{F} is a vector-valued function class with the range $[a, b]$. For any $\boldsymbol{\xi} \geq \mathbf{0}$ such that

$$N \geq \frac{8\Gamma(\mathbf{\Lambda})}{1 - 2\Upsilon(\mathbf{\Lambda})}, \quad (34)$$

then there holds that

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}'_N \mathbf{f} - \mathbf{E}_N \mathbf{f}| > \frac{\boldsymbol{\xi}}{2} \right\}, \quad (35)$$

where

$$\Gamma(\mathbf{\Lambda}) := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \frac{m(b-a)^2}{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2}, \quad \Upsilon(\mathbf{\Lambda}) := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}),$$

$\mathbf{\Lambda} = \{1, \dots, M\}$ is an index set and $\Lambda^{[m]}$ is a subset of $\mathbf{\Lambda}$ with the cardinality of m .

The following is the symmetrization result incorporating the covariance information between every two tasks.

Theorem B.2 Assume that \mathcal{F} is a vector-valued function class with the range $[a, b]$. For any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > 0$ such that

$$N \geq \frac{8\Gamma_2}{1 - 2(\Upsilon(\Lambda) + \Upsilon_2)}, \quad (36)$$

then there holds that

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} \{ |\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}| \} > \boldsymbol{\xi} \right\} \leq 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} \{ |\mathbf{E}'_N\mathbf{f} - \mathbf{E}_N\mathbf{f}| \} > \frac{\boldsymbol{\xi}}{2} \right\}, \quad (37)$$

where

$$\Gamma_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{m(b-a)^2}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2}, \quad \Upsilon_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{8 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}_{\mathcal{F}}(i_1, i_2)}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2}$$

and $\text{Cov}_{\mathcal{F}}(i_1, i_2)$ is defined as

$$\text{Cov}_{\mathcal{F}}(i, j) := \max_{(f^{[1]}, \dots, f^{[M]})^T \in \mathcal{F}} \text{Cov}(f^{[i]}(\mathbf{z}^{[i]}), f^{[j]}(\mathbf{z}^{[j]})) \quad (38)$$

with $\mathbf{z}^{[i]}$ and $\mathbf{z}^{[j]}$ ($1 \leq i, j \leq M$) being the random variables of the tasks $\mathcal{Z}^{[i]}$ and $\mathcal{Z}^{[j]}$, respectively.

C Proofs of Main Results

C.1 Proof of Lemma A.1

Proof of Lemma A.1. Let $\mathbf{t} = \left| \frac{1}{N} \sum_{n=1}^N \mathbf{s}_n - \boldsymbol{\mu} \right|$. The event $|\mathbf{t}| > \boldsymbol{\xi}$ contains 2^M possibilities: for any $1 \leq m \leq M$, there are m components of the vector \mathbf{t} such that $t^{[i_k]} > \xi^{[i_k]}$ ($1 \leq k \leq m$) and the rest are of the case that $t^{[i_k]} < -\xi^{[i_k]}$ ($1 \leq k \leq M - m$). For convenience, we also denote $\{\mathcal{P}_i\}_{i=1}^{2^M}$ as the collection of all 2^M possibilities.

According to Theorem 1 in [9], the following result is valid for any possibility \mathcal{P}_i ($1 \leq i \leq 2^M$):

$$\Pr \{ \mathcal{P}_i \} \leq \prod_{m=0}^M \left(\frac{\mu^{[m]}}{p^{[m]}} \right)^{p^{[m]}N}, \quad (39)$$

where $p^{[m]} = \mu^{[m]} + \xi^{[m]}$ ($m = 1, \dots, M$), $\mu_0 = 1 - \sum_{m=1}^M \mu^{[m]}$ and $p_0 = 1 - \sum_{m=1}^M p^{[m]}$. Then, we have

$$\Pr \left\{ \left| \frac{1}{N} \sum_{n=1}^N \mathbf{s}_n - \boldsymbol{\mu} \right| > \boldsymbol{\xi} \right\} \leq 2^M \prod_{m=0}^M \left(\frac{\mu^{[m]}}{p^{[m]}} \right)^{p^{[m]}N}. \quad (40)$$

Then, consider

$$\begin{aligned}
\prod_{m=0}^M \left(\frac{\mu^{[m]}}{p^{[m]}} \right)^{p^{[m]}N} &= \exp \left\{ N \sum_{m=0}^M p^{[m]} \log \left(\frac{\mu^{[m]}}{p^{[m]}} \right) \right\} \\
&= \exp \left\{ N \left(\left(1 - \sum_{m=1}^M p^{[m]} \right) \log \left(\frac{1 - \sum_{m=1}^M \mu^{[m]}}{1 - \sum_{m=1}^M p^{[m]}} \right) + \sum_{m=1}^M p^{[m]} \log \left(\frac{\mu^{[m]}}{p^{[m]}} \right) \right) \right\} \\
&\leq \exp \left\{ N \left(\sum_{m=1}^M \left(1 - p^{[m]} \right) \log \left(\frac{1 - \mu^{[m]}}{1 - p^{[m]}} \right) + \sum_{m=1}^M p^{[m]} \log \left(\frac{\mu^{[m]}}{p^{[m]}} \right) \right) \right\} \quad (*) \\
&= \exp \left\{ -N \sum_{m=1}^M \int_{\mu^{[m]}}^{p^{[m]}} \left(\frac{p^{[m]}}{x} - \frac{1 - p^{[m]}}{1 - x} \right) dx \right\} \\
&= \exp \left\{ -N \sum_{m=1}^M \int_{\mu^{[m]}}^{p^{[m]}} \frac{p^{[m]} - x}{x(1 - x)} dx \right\} \\
&\leq \exp \left\{ -N \sum_{m=1}^M 4 \int_{\mu^{[m]}}^{p^{[m]}} (p^{[m]} - x) dx \right\} \\
&= \exp \left\{ -2N \sum_{m=1}^M (p^{[m]} - \mu^{[m]})^2 \right\} = \exp \left\{ -2N \sum_{m=1}^M (\xi^{[m]})^2 \right\}, \tag{41}
\end{aligned}$$

because $x(1-x) \leq 1/4$ for any $x \in \mathbb{R}$, and the step $(*)$ is followed from the fact that the function f is subadditive if f is concave and $f(0) \geq 0$. \blacksquare

C.2 Proof of Lemma B.1

Proof of Lemma B.1. Given M tasks $\mathcal{Z}^{(1)}, \dots, \mathcal{Z}^{[M]}$ and a vector-valued function class \mathcal{F} , let $\mathbf{\Lambda} := \{1, \dots, M\}$ be an index set and $\Lambda^{[m]}$ be a subset of $\mathbf{\Lambda}$ with the cardinality of m . For any $\Lambda^{[m]} \subset \mathbf{\Lambda}$ and any $\boldsymbol{\xi} = (\xi^{(1)}, \dots, \xi^{[M]})^T > \mathbf{0}$, define

$$\psi(\Lambda^{[m]}, \boldsymbol{\xi}) := \Pr\{\{s^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}} \mid \{s^{[i]} \leq \xi^{[i]}\}_{i \in \overline{\Lambda^{[m]}}}\} - \Pr\{\{s^{[i]} > \xi^{[i]}\}_{i \in \Lambda^{[m]}}\}. \tag{42}$$

Then, the event $\mathbf{s} \not\leq \boldsymbol{\xi}$ contains the following possibilities:

- $\mathcal{P}^{[1]}$: there is only one index $\{i\} = \Lambda^{[1]} \subset \mathbf{\Lambda}$ satisfying that $s^{[i]} > \xi^{[i]}$;
- $\mathcal{P}^{[m]}$: there are only m ($1 < m < M$) indices $\{i^{[1]}, \dots, i^{[m]}\} = \Lambda^{[m]} \subset \mathbf{\Lambda}$ satisfying that $s^{[i_k]} > \xi^{[i_k]}$ ($1 \leq k \leq m$);
- $\mathcal{P}^{[M]}$: $s^{[m]} > \xi^{[m]}$ holds for any $1 \leq m \leq M$.

Thus, we have

$$\Pr\{\mathbf{s} \not\leq \boldsymbol{\xi}\} = \Pr\{\mathcal{P}^{[1]}\} + \dots + \Pr\{\mathcal{P}^{[M]}\}. \tag{43}$$

According to Chebyshev's inequality and (42), we have

$$\Pr\{\mathcal{P}^{[1]}\} = \sum_{m=1}^M \left(\psi(\{m\}, \boldsymbol{\xi}) + \Pr\{s^{[m]} > \xi^{[m]}\} \right) \leq \sum_{m=1}^M \left(\psi(\{m\}, \boldsymbol{\xi}) + \frac{\mathbb{E}\{(s^{[m]})^2\}}{(\xi^{[m]})^2} \right), \tag{44}$$

and for any $2 \leq m \leq M$,

$$\begin{aligned}
\Pr\{\mathcal{P}^{[m]}\} &= \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} (\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \Pr\{s^{[i]} > \xi^{[i]} : i \in \Lambda^{[m]}\}) \\
&= \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} (\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \Pr\{(s^{[i]})^2 > (\xi^{[i]})^2 : i \in \Lambda^{[m]}\}) \\
&\leq \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \Pr \left\{ \sqrt{\sum_{i \in \Lambda^{[m]}} (s^{[i]})^2} > \sqrt{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right\} \right) \\
&\leq \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\mathbb{E} \left\{ \sum_{i \in \Lambda^{[m]}} (s^{[i]})^2 \right\}}{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right) \\
&= \sum_{\Lambda^{[m]} \subset \mathbf{\Lambda}} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\sum_{i \in \Lambda^{[m]}} \mathbb{E}\{s^{[i]}\}^2}{\sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right). \tag{45}
\end{aligned}$$

The combination of (43), (44) and (45) leads to the result (32). This completes the proof. \blacksquare

C.3 Proof of Lemma B.2

Proof of Lemma B.2. The event $\mathbf{s} \not\leq \boldsymbol{\xi}$ contains the following possibilities:

- $\mathcal{P}^{[1]}$: there is only one index $\{i\} = \Lambda^{[1]} \subset \mathbf{\Lambda}$ satisfying that $s^{[i]} > \xi^{[i]}$;
- $\mathcal{P}^{[m]}$: there are only m ($1 < m < M$) indices $\{i^{[1]}, \dots, i^{[m]}\} = \Lambda^{[m]} \subset \mathbf{\Lambda}$ satisfying that $s^{[i_k]} > \xi^{[i_k]}$ ($1 \leq k \leq m$);
- $\mathcal{P}^{[M]}$: $s^{[m]} > \xi^{[m]}$ holds for any $1 \leq m \leq M$.

Thus, we have

$$\Pr\{\mathbf{s} \not\leq \boldsymbol{\xi}\} = \Pr\{\mathcal{P}^{[1]}\} + \dots + \Pr\{\mathcal{P}^{[M]}\}. \tag{46}$$

According to Chebyshev's inequality and (42), we have

$$\Pr\{\mathcal{P}^{[1]}\} = \sum_{m=1}^M (\psi(\{m\}, \boldsymbol{\xi}) + \Pr\{s^{[m]} > \xi^{[m]}\}) \leq \sum_{m=1}^M \left(\psi(\{m\}, \boldsymbol{\xi}) + \frac{\mathbb{E}\{(s^{[m]})^2\}}{(\xi^{[m]})^2} \right), \tag{47}$$

and for any $2 \leq m \leq M$,

$$\begin{aligned}
\Pr\{\mathcal{P}^{[m]}\} &= \sum_{\Lambda^{[m]} \subset \Lambda} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \Pr\{s^{[i]} > \xi^{[i]} : i \in \Lambda^{[m]}\} \right) \\
&\leq \sum_{\Lambda^{[m]} \subset \Lambda} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \Pr\left\{ \sum_{i \in \Lambda^{[m]}} (s^{[i]}) > \sum_{i \in \Lambda^{[m]}} (\xi^{[i]}) \right\} \right) \\
&\leq \sum_{\Lambda^{[m]} \subset \Lambda} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\mathbb{E}\left\{ \left(\sum_{i \in \Lambda^{[m]}} s^{[i]} \right)^2 \right\}}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right) \\
&= \sum_{\Lambda^{[m]} \subset \Lambda} \left(\psi(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\sum_{i \in \Lambda^{[m]}} \mathbb{E}\{(s^{[i]})^2\} + 2 \sum_{\substack{i, j \in \Lambda^{[m]} \\ i < j}} \mathbb{E}\{s^{[i]} s^{[j]}\}}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right). \tag{48}
\end{aligned}$$

The combination of (46), (47) and (48) leads to the result (33). This completes the proof. \blacksquare

C.4 Proof of Theorem B.1

Proof of Theorem B.1. Let $\mathbf{f}_N = (\widehat{f}^{[1]}, \dots, \widehat{f}^{[M]})^T$ be the vector-valued function achieving the supremum

$$\sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}|.$$

According to the triangle inequality, we have

$$|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| - |\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}\mathbf{f}_N| \leq |\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N|, \tag{49}$$

and thus

$$\begin{aligned}
\mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \boldsymbol{\xi}\}} \mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N \mathbf{f}_N| \leq \boldsymbol{\xi}/2\}} &= \mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \boldsymbol{\xi}\} \wedge \{|\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}\mathbf{f}_N| \leq \boldsymbol{\xi}/2\}} \\
&\leq \mathbf{1}_{\{|\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \boldsymbol{\xi}/2\}}. \tag{50}
\end{aligned}$$

Taking expectations with respect to the ghost samples gives

$$\mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \boldsymbol{\xi}\}} \Pr' \left\{ |\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N \mathbf{f}_N| \leq \frac{\boldsymbol{\xi}}{2} \right\} \leq \Pr' \left\{ |\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \frac{\boldsymbol{\xi}}{2} \right\}. \tag{51}$$

According to Lemma B.1, since the samples $\mathbf{z}_n^{[m]}$ ($1 \leq m \leq M$, $1 \leq n \leq M$) are independent of

each other, we have

$$\begin{aligned}
& \Pr' \left\{ |\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N \mathbf{f}_N| \not\leq \frac{\boldsymbol{\xi}}{2} \right\} \\
&= \Pr \left\{ \begin{pmatrix} \left| \sum_{n=1}^N (\mathbf{E}^{[1]} \hat{\mathbf{f}}^{[1]} - \hat{\mathbf{f}}^{[1]}(\mathbf{z}_n^{[1]})) \right| \\ \vdots \\ \left| \sum_{n=1}^N (\mathbf{E}^{[M]} \hat{\mathbf{f}}^{[M]} - \hat{\mathbf{f}}^{[M]}(\mathbf{z}_n^{[M]})) \right| \end{pmatrix} \not\leq \begin{pmatrix} \frac{N\xi^{[1]}}{2} \\ \vdots \\ \frac{N\xi^{[M]}}{2} \end{pmatrix} \right\} \\
&\leq \Pr \left\{ \begin{pmatrix} \sum_{n=1}^N |\mathbf{E}^{[1]} \hat{\mathbf{f}}^{[1]} - \hat{\mathbf{f}}^{[1]}(\mathbf{z}_n^{[1]})| \\ \vdots \\ \sum_{n=1}^N |\mathbf{E}^{[M]} \hat{\mathbf{f}}^{[M]} - \hat{\mathbf{f}}^{[M]}(\mathbf{z}_n^{[M]})| \end{pmatrix} \not\leq \begin{pmatrix} \frac{N\xi^{[1]}}{2} \\ \vdots \\ \frac{N\xi^{[M]}}{2} \end{pmatrix} \right\} \\
&\leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{N \sum_{i \in \Lambda^{[m]}} \mathbf{E}^{[i]} \left\{ (\mathbf{E}^{[i]} \hat{\mathbf{f}}^{[i]} - \hat{\mathbf{f}}^{[i]}(\mathbf{z}^{[i]})^2 \right\}}{\frac{N^2}{4} \sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right) \\
&= \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{\sum_{i \in \Lambda^{[m]}} 4\text{Var}^{[i]}(\hat{\mathbf{f}}^{[i]}(\mathbf{z}^{[i]}))}{N \sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right) \quad (*) \\
&\leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{4m(b-a)^2}{N \sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right), \tag{52}
\end{aligned}$$

where the step (*) is followed from the fact that for each task $\mathcal{Z}^{[m]}$ ($1 \leq m \leq M$), the samples $\{z_n^{[m]}\}_{n=1}^N$ are independent.

Hence, we get

$$\begin{aligned}
& \mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N \mathbf{f}_N| > \boldsymbol{\xi}\}} \left(1 - \left(\sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{4m(b-a)^2}{N \sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right) \right) \\
&\leq \Pr' \left\{ |\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \frac{\boldsymbol{\xi}}{2} \right\}. \tag{53}
\end{aligned}$$

Taking the expectation with respect to the sample collection $\{\mathbf{Z}_N^{[m]}\}_{m=1}^M$ of the tasks $\mathcal{Z}^{[1]}, \dots, \mathcal{Z}^{[M]}$ and letting

$$\sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \boldsymbol{\xi}) + \frac{4m(b-a)^2}{N \sum_{i \in \Lambda^{[m]}} (\xi^{[i]})^2} \right) \leq \frac{1}{2}, \tag{54}$$

we then have for any $\boldsymbol{\xi} > \mathbf{0}$,

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}'_N \mathbf{f} - \mathbf{E}_N \mathbf{f}| > \frac{\boldsymbol{\xi}}{2} \right\}.$$

This completes the proof. ■

C.5 Proof of Theorem B.2

Proof of Theorem B.2. Let $\mathbf{f}_N = (\hat{f}_1, \dots, \hat{f}_M)^T$ be the vector-valued function achieving the supremum

$$\sup_{\mathbf{f} \in \mathcal{F}} \{|\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}|\}.$$

Similar to the proof of Theorem B.1, we have

$$\mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N\mathbf{f}_N| > \xi\}} \Pr' \left\{ |\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N\mathbf{f}_N| \leq \frac{\xi}{2} \right\} \leq \Pr' \left\{ |\mathbf{E}'_N\mathbf{f}_N - \mathbf{E}_N\mathbf{f}_N| > \frac{\xi}{2} \right\}. \quad (55)$$

According to Lemma B.2, we have

$$\begin{aligned} & \Pr' \left\{ |\mathbf{E}\mathbf{f}_N - \mathbf{E}'_N\mathbf{f}_N| \not\leq \frac{\xi}{2} \right\} \\ & \leq \Pr \left\{ \begin{pmatrix} \sum_{n=1}^N |\mathbf{E}\hat{f}_1 - \hat{f}_1(\mathbf{z}_n^{[1]})| \\ \vdots \\ \sum_{n=1}^N |\mathbf{E}\hat{f}_M - \hat{f}_M(\mathbf{z}_n^{[M]})| \end{pmatrix} \not\leq \begin{pmatrix} N\xi^{[1]}/2 \\ \vdots \\ N\xi^{[M]}/2 \end{pmatrix} \right\} \\ & \leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \xi) + \frac{N \sum_{i \in \Lambda^{[m]}} \text{Var}(\hat{f}^{[i]}(\mathbf{z}^{[i]})) + 2N^2 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}(\hat{f}^{(i_1)}(\mathbf{z}^{(i_1)}), \hat{f}^{(i_2)}(\mathbf{z}^{(i_2)}))}{\frac{N^2}{4} \left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right) \\ & = \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \xi) + \frac{4 \sum_{i \in \Lambda^{[m]}} \text{Var}(\hat{f}^{[i]}(\mathbf{z}^{[i]}))}{N \left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} + \frac{8 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}(\hat{f}^{(i_1)}(\mathbf{z}^{(i_1)}), \hat{f}^{(i_2)}(\mathbf{z}^{(i_2)}))}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right) \\ & \leq \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \left(\phi_{\mathcal{F}}(\Lambda^{[m]}, \xi) + \frac{4m(b-a)^2}{N \left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} + \frac{8 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}_{\mathcal{F}}(i_1, i_2)}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2} \right) \quad (56) \end{aligned}$$

Moreover, define

$$\Gamma_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{m(b-a)^2}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2},$$

and

$$\Upsilon_2 := \sum_{m=1}^M \sum_{\Lambda^{[m]} \subset \Lambda} \frac{8 \sum_{\substack{i_1 < i_2 \\ i_1, i_2 \in \Lambda^{[m]}}} \text{Cov}_{\mathcal{F}}(i_1, i_2)}{\left(\sum_{i \in \Lambda^{[m]}} \xi^{[i]} \right)^2}.$$

Hence, we get

$$\mathbf{1}_{\{|\mathbf{E}\mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \xi\}} \left(1 - \left(\frac{4\Gamma_2}{N} + \Upsilon(\Lambda) + \Upsilon_2 \right) \right) \leq \Pr' \left\{ |\mathbf{E}'_N \mathbf{f}_N - \mathbf{E}_N \mathbf{f}_N| > \frac{\xi}{2} \right\}. \quad (57)$$

Taking the expectation with respect to $\{\mathbf{Z}_N^{[m]}\}_{m=1}^M$ and letting

$$\frac{4\Gamma_2}{N} + \Upsilon(\Lambda) + \Upsilon_2 \leq \frac{1}{2}, \quad (58)$$

we then have for any $\xi > 0$

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| > \xi \right\} \leq 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}} |\mathbf{E}'_N \mathbf{f} - \mathbf{E}_N \mathbf{f}| > \frac{\xi}{2} \right\}.$$

This completes the proof. ■

C.6 Proof of Theorem 5.1

Proof of Theorem 5.1. For any $1 \leq m \leq M$, consider $\{\epsilon_n^{[m]}\}_{n=1}^N$ as independent Rademacher random variables, *i.e.*, independent $\{\pm 1\}$ -valued random variables with equal probability of taking either value. Given an $\{\epsilon_n^{[m]}\}_{n=1}^N$ and a $\mathbf{Z}_{2N}^{[m]}$, denote

$$\vec{\epsilon}^{[m]} := (\epsilon_1^{[m]}, \dots, \epsilon_N^{[m]}, -\epsilon_1^{[m]}, \dots, -\epsilon_N^{[m]})^T \in \{\pm 1\}^{2N}, \quad 1 \leq m \leq M, \quad (59)$$

and for any $\mathbf{f} = (f_1, \dots, f_M)^T \in \mathcal{F}_c^{\mathbf{R}}$,

$$\vec{f}^{[m]}(\mathbf{Z}_{2N}^{[m]}) := (f^{[m]}(\mathbf{z}_1^{[m]}), \dots, f^{[m]}(\mathbf{z}_N^{[m]}), f^{[m]}(\mathbf{z}_1^{[m]}), \dots, f^{[m]}(\mathbf{z}_N^{[m]}))^T \in [a, b]^{2N}. \quad (60)$$

According to Theorem B.1, given any $\xi > 0$ and for any $N \in \mathbb{N}$ satisfying Condition (34), we have

$$\begin{aligned} & \Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} |\mathbf{E}\mathbf{f} - \mathbf{E}_N \mathbf{f}| > \xi \right\} \\ & \leq 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} |\mathbf{E}'_N \mathbf{f} - \mathbf{E}_N \mathbf{f}| > \frac{\xi}{2} \right\} \quad (\text{by Theorem B.2}) \\ & = 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} \left(\begin{array}{c} \left| \frac{1}{N} \sum_{n=1}^N (f(\mathbf{z}'_n^{[1]}) - f(\mathbf{z}_n^{[1]})) \right| \\ \vdots \\ \left| \frac{1}{N} \sum_{n=1}^N (f(\mathbf{z}'_n^{[M]}) - f(\mathbf{z}_n^{[M]})) \right| \end{array} \right) > \left(\begin{array}{c} \frac{\xi^{[1]}}{2} \\ \vdots \\ \frac{\xi^{[M]}}{2} \end{array} \right) \right\} \\ & = 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} \left(\begin{array}{c} \left| \frac{1}{N} \sum_{n=1}^N \epsilon_n^{[1]} (f(\mathbf{z}'_n^{[1]}) - f(\mathbf{z}_n^{[1]})) \right| \\ \vdots \\ \left| \frac{1}{N} \sum_{n=1}^N \epsilon_n^{[M]} (f(\mathbf{z}'_n^{[M]}) - f(\mathbf{z}_n^{[M]})) \right| \end{array} \right) > \left(\begin{array}{c} \frac{\xi^{[1]}}{2} \\ \vdots \\ \frac{\xi^{[M]}}{2} \end{array} \right) \right\} \\ & = 2\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^{\mathbf{R}}} \left(\begin{array}{c} \left| \frac{1}{2N} \langle \vec{\epsilon}^{[1]}, \vec{f}_1(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{2N} \langle \vec{\epsilon}^{[M]}, \vec{f}_M(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{array} \right) > \left(\begin{array}{c} \frac{\xi^{[1]}}{4} \\ \vdots \\ \frac{\xi^{[M]}}{4} \end{array} \right) \right\}. \end{aligned} \quad (61)$$

For any given sample collection $\{\mathbf{Z}_{2N}^{[m]}\}_{m=1}^M$ of the tasks $\mathcal{Z}^{[1]}, \dots, \mathcal{Z}^{[M]}$, let $\Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)$ be the cover of \mathcal{F}_c^R w.r.t. the radius-vectors $\boldsymbol{\xi}/8$. Since \mathcal{F}_c^R is composed of the functions with the range $[a, b]$, we assume that the same holds for any $\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)$. If $\mathbf{f}_\dagger = (f_\dagger^{[1]}, \dots, f_\dagger^{[M]})^T$ is a vector-valued function that achieves

$$\sup_{\mathbf{f} \in \mathcal{F}_c^R} \left(\begin{array}{c} \left| \frac{1}{2N} \langle \vec{\epsilon}^{[1]}, \vec{f}^{[1]}(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{2N} \langle \vec{\epsilon}^{[M]}, \vec{f}^{[M]}(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{array} \right) > \begin{pmatrix} \frac{\xi^{[1]}}{4} \\ \vdots \\ \frac{\xi^{[M]}}{4} \end{pmatrix},$$

there must be an $\mathbf{h}_\dagger = (h_\dagger^{[1]}, \dots, h_\dagger^{[M]})^T \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)$ such that, for any $1 \leq m \leq M$,

$$\frac{1}{2N} \sum_{n=1}^N \left(|f_\dagger^{[m]}(\mathbf{z}_n^{[m]}) - h_\dagger^{[m]}(\mathbf{z}_n^{[m]})| + |f_\dagger^{[m]}(\mathbf{z}_n^{[m]}) - h_\dagger^{[m]}(\mathbf{z}_n^{[m]})| \right) < \frac{\xi^{[m]}}{8},$$

and meanwhile,

$$\left| \frac{1}{2N} \langle \vec{\epsilon}^{[m]}, \vec{h}_\dagger^{[M]}(\mathbf{Z}_{2N}^{[m]}) \rangle \right| > \frac{\xi^{[m]}}{8}.$$

Therefore, we arrive at

$$\begin{aligned} & \Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} \left(\begin{array}{c} \left| \frac{1}{2N} \langle \vec{\epsilon}^{[1]}, \vec{f}^{[1]}(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{2N} \langle \vec{\epsilon}^{[M]}, \vec{f}^{[M]}(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{array} \right) > \begin{pmatrix} \frac{\xi^{[1]}}{4} \\ \vdots \\ \frac{\xi^{[M]}}{4} \end{pmatrix} \right\} \\ & \leq \Pr \left\{ \sup_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} \left(\begin{array}{c} \left| \frac{1}{2N} \langle \vec{\epsilon}^{[1]}, \vec{h}^{[1]}(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{2N} \langle \vec{\epsilon}^{[M]}, \vec{h}^{[M]}(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{array} \right) > \begin{pmatrix} \frac{\xi^{[1]}}{8} \\ \vdots \\ \frac{\xi^{[M]}}{8} \end{pmatrix} \right\}. \end{aligned} \quad (62)$$

On the other hand, given a $\boldsymbol{\xi} > \mathbf{0}$ and for any $N \in \mathbb{N}$ satisfying Condition (34),

$$\begin{aligned}
& \Pr \left\{ \sup_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} \begin{pmatrix} \left| \frac{1}{2N} \langle \vec{\epsilon}^{[1]}, \vec{h}^{[1]}(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{2N} \langle \vec{\epsilon}^{[M]}, \vec{h}^{[M]}(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{pmatrix} > \begin{pmatrix} \frac{\xi^{[1]}}{8} \\ \vdots \\ \frac{\xi^{[M]}}{8} \end{pmatrix} \right\} \\
&= \Pr \left\{ \sup_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} \begin{pmatrix} \left| \frac{1}{N} \langle \vec{\epsilon}^{[1]}, \vec{h}^{[1]}(\mathbf{Z}_{2N}^{[1]}) \rangle \right| \\ \vdots \\ \left| \frac{1}{N} \langle \vec{\epsilon}^{[M]}, \vec{h}^{[M]}(\mathbf{Z}_{2N}^{[M]}) \rangle \right| \end{pmatrix} > \begin{pmatrix} \frac{\xi^{[1]}}{4} \\ \vdots \\ \frac{\xi^{[M]}}{4} \end{pmatrix} \right\} \\
&= \Pr \left\{ \sup_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} |\mathbf{E}'_N \mathbf{h} - \mathbf{E}_N \mathbf{h}| > \frac{\boldsymbol{\xi}}{4} \right\} \quad (\text{similar to (61)}) \\
&\leq \Pr \left\{ \sum_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} |\mathbf{E}'_N \mathbf{h} - \mathbf{E}_N \mathbf{h}| > \frac{\boldsymbol{\xi}}{4} \right\} \\
&\leq \Pr \left\{ \sum_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} |\mathbf{E} \mathbf{h} - \mathbf{E}_N \mathbf{h}| + |\mathbf{E} \mathbf{h} - \mathbf{E}'_N \mathbf{h}| > \frac{\boldsymbol{\xi}}{4} \right\} \\
&\leq 2 \Pr \left\{ \sum_{\mathbf{h} \in \Omega_{p,N}(\mathcal{F}_c^R, \boldsymbol{\xi}/8)} |\mathbf{E} \mathbf{h} - \mathbf{E}_N \mathbf{h}| > \frac{\boldsymbol{\xi}}{8} \right\} \\
&\leq 2^{M+1} \mathcal{N}_1(\mathcal{F}_c^R, \boldsymbol{\xi}/8, 2N) \exp \left\{ \frac{-N \sum_{m=1}^M (\xi^{[m]})^2}{32M^2(b-a)^2} \right\}. \tag{63}
\end{aligned}$$

The last inequality of (63) is derived from Definition (4.2) and Theorem A.1.

The combination of (61), (62) and (63) leads to the result: given any $\boldsymbol{\xi} > \mathbf{0}$, there holds that for any $N \in \mathbb{N}$ satisfying Condition (34),

$$\Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E} \mathbf{f} - \mathbf{E}_N \mathbf{f}| > \boldsymbol{\xi} \right\} \leq 2^{M+2} \mathcal{N}_1(\mathcal{F}_c^R, \boldsymbol{\xi}/8, 2N) \exp \left\{ \frac{-N \sum_{m=1}^M (\xi^{[m]})^2}{32M^2(b-a)^2} \right\}.$$

This completes the proof. ■

C.7 Proof of Theorem 5.2

Before the formal proof, we present a necessary lemma.

Lemma C.1 *Let $\mathbf{s}_n = (s_n^{[1]}, \dots, s_n^{[M]}) \in \mathbb{R}^M$ ($1 \leq n \leq N$) be N i.i.d. random vectors. Then, there holds that for any $\boldsymbol{\xi} = (\xi^{[1]}, \dots, \xi^{[M]})^T > \mathbf{0}$,*

$$\Pr \left\{ \sum_{n=1}^N \mathbf{s}_n \leq N \boldsymbol{\xi} \right\} \leq 2^M \Pr \{ \mathbf{s}_1 \leq 2 \boldsymbol{\xi} \}. \tag{64}$$

Proof. For any $1 \leq m \leq M$, we have

$$\sum_{n=1}^N s_n^{[m]} \geq \sum_{n=1}^N s_n^{[m]} \mathbf{1}_{\{s_n^{[m]} > 2\xi^{[m]}\}} \geq 2\xi^{[m]} \sum_{n=1}^N \mathbf{1}_{\{s_n^{[m]} > 2\xi^{[m]}\}}.$$

Hence, it is followed from the conditional Markov inequality that

$$\begin{aligned} \Pr \left\{ \sum_{n=1}^N \mathbf{s}_n \leq N\boldsymbol{\xi} \right\} &\leq \Pr \left\{ \begin{pmatrix} 2\xi^{[1]} \sum_{n=1}^N \mathbf{1}_{\{s_n^{[1]} > 2\xi^{[1]}\}} \\ \vdots \\ 2\xi^{[M]} \sum_{n=1}^N \mathbf{1}_{\{s_n^{[M]} > 2\xi^{[M]}\}} \end{pmatrix} \leq N \begin{pmatrix} \xi^{[1]} \\ \vdots \\ \xi^{[M]} \end{pmatrix} \right\} \\ &= \Pr \left\{ \begin{pmatrix} \sum_{n=1}^N \mathbf{1}_{\{s_n^{[1]} \leq 2\xi^{[1]}\}} \\ \vdots \\ \sum_{n=1}^N \mathbf{1}_{\{s_n^{[M]} \leq 2\xi^{[M]}\}} \end{pmatrix} \geq (1 - 2^{-1})N \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right\} \\ &= \Pr \left\{ \sum_{n=1}^N \mathbf{1}_{\{s_n^{[1]} \leq 2\xi^{[1]}\}} \geq 2^{-1}N \mid \mathcal{A}_2^M \right\} \Pr \{ \mathcal{A}_2^M \} \\ &\leq \frac{\mathbb{E} \left\{ \sum_{n=1}^N \mathbf{1}_{\{s_n^{[1]} \leq 2\xi^{[1]}\}} \mid \mathcal{A}_2^M \right\}}{2^{-1}N} \Pr \{ \mathcal{A}_2^M \} \\ &\leq \frac{N \Pr \{ s_1^{[1]} \leq 2\xi^{[1]} \mid \mathcal{A}_2^M \}}{2^{-1}N} \Pr \{ \mathcal{A}_2^M \} = 2 \Pr \{ s_1^{[1]} \leq 2\xi^{[1]}, \mathcal{A}_2^M \}, \end{aligned}$$

where \mathcal{A}_2^M stands for the event that $\left\{ \sum_{n=1}^N \mathbf{1}_{\{s_n^{[m]} \leq 2\xi^{[m]}\}} \right\}_{m=2}^M$. Then, following this way, we have

$$\begin{aligned} \Pr \left\{ \sum_{n=1}^N \mathbf{s}_n \leq N\boldsymbol{\xi} \right\} &\leq 2 \Pr \{ s_1^{[1]} \leq 2\xi^{[1]}, \mathcal{A}_2^M \} \leq 2^2 \Pr \{ s_1^{[1]} \leq 2\xi^{[1]}, s_1^{[2]} \leq 2\xi^{[2]}, \mathcal{A}_3^M \} \\ &\leq \cdots \leq 2^M \Pr \{ s_1^{[1]} \leq 2\xi^{[1]}, s_1^{[2]} \leq 2\xi^{[2]}, \dots, s_1^{[M]} \leq 2\xi^{[M]} \} = 2^M \Pr \{ \mathbf{s}_1 \leq 2\boldsymbol{\xi} \}. \end{aligned}$$

This completes the proof. ■

Next, we come up with the proof of Theorem 5.2.

Proof of Theorem 5.2. Let $\hat{\mathbf{f}}_* = (f_*^{[1]}, \dots, f_*^{[M]})^T$ be the vector-valued function achieving the supremum $\sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}|$. Then, it is followed from Lemma C.1 that

$$\begin{aligned} \Pr \left\{ \sup_{\mathbf{f} \in \mathcal{F}_c^R} |\mathbf{E}\mathbf{f} - \mathbf{E}_N\mathbf{f}| \leq \boldsymbol{\xi} \right\} &= \Pr \{ |\mathbf{E}\hat{\mathbf{f}}_* - \mathbf{E}_N\hat{\mathbf{f}}_*| \leq \boldsymbol{\xi} \} \leq 2^M \Pr \{ \mathbf{s}_* \leq 2\boldsymbol{\xi} \} \\ &\leq 2^M \sup_{\mathbf{f} \in \mathcal{F}_c^R} \Pr \{ \mathbf{s} \leq 2\boldsymbol{\xi} \}, \end{aligned}$$

where $\mathbf{s}_* = (s_*^{[1]}, \dots, s_*^{[M]})^T$ with $s_*^{[m]} := |\mathbb{E}^{[m]} f_*^{[m]} - f_*^{[m]}(\mathbf{z}^{[m]})|$ for any $1 \leq m \leq M$. ■

C.8 Proof of Theorem 5.3

Proof of Theorem 5.3. Denote $\mathbf{t}_N = (t^{[1]}, \dots, t^{[M]})^T$ with $t^{[i]} := |E^{[i]}f^{[i]} - E_N^{[i]}f^{[i]}| > \xi^{[i]}$. The event $\mathbf{t}_N \not\leq \boldsymbol{\xi}$ contains the following possibilities:

- $\mathcal{P}^{[1]}$: there is only one index $\{i\} = \Lambda^{[1]} \subset \mathbf{\Lambda}$ satisfying that $t^{[i]} > \xi^{[i]}$;
- $\mathcal{P}^{[m]}$: there are m ($1 < m < M$) indices $\{i^{[1]}, \dots, i^{[m]}\} = \Lambda^{[m]} \subset \mathbf{\Lambda}$ satisfying that $t^{[i_k]} > \xi^{[i_k]}$ ($1 \leq k \leq m$);
- $\mathcal{P}^{[M]}$: $t^{[m]} > \xi^{[m]}$ holds for any $1 \leq m \leq M$.

Thus, we have

$$\Pr\{\mathbf{t}_N \not\leq \boldsymbol{\xi}\} = \Pr\{\mathcal{P}^{[1]}\} + \dots + \Pr\{\mathcal{P}^{[M]}\}. \quad (65)$$

Then, the combination of Definition 3.2, Theorems 5.1&5.2 and (65) leads to the result (17). Moreover, since $\Pr\left\{\left\{s^{[\lambda]} \leq 2\xi^{[\lambda]}\right\}_{\lambda \in \overline{\Lambda^{[m]}}}\right\} \leq 1$ holds for any $\Lambda^{[m]} \subset \mathbf{\Lambda}$, the result (19) can be directly obtained. This completes the proof. \blacksquare